ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

The Relationship Between Single-Word Speech Severity and Intelligibility in Childhood Apraxia of Speech

Karen V. Chenausky,[a,b,c] Danielle Gagné,[a] Kaila L. Stipancic,[a,d] Aaron Shield,[e] and Jordan R. Green[a,f]

[a] MGH Institute of Health Professions, Boston, MA [b] Harvard Medical School, Boston, MA [c] Department of Psychological and Brain Sciences, Boston University, MA [d] Department of Communicative Disorders and Sciences, University at Buffalo, NY [e] Department of Speech Pathology & Audiology, Miami University, Oxford, OH [f] Speech and Hearing and Biosciences and Technology Program, Harvard University, Boston, MA

ARTICLE INFO

Article History:
Received April 13, 2021
Revision received July 29, 2021
Accepted October 24, 2021

Editor-in-Chief: Bharath Chandrasekaran
Editor: Julie D. Anderson

https://doi.org/10.1044/2021_JSLHR-21-00213

ABSTRACT

Purpose: The purpose of this study was to investigate the association between perceived single-word speech severity and intelligibility in children with childhood apraxia of speech (CAS), with and without comorbid language impairment (LI), and to investigate the contribution of different CAS signs to perceived single-word speech severity and single-word intelligibility.

Method: Thirty children with CAS, 18 with comorbid LI, completed the Goldman-Fristoe Test of Articulation–Second Edition (GFTA-2). Trained judges coded children's responses for signs of CAS and percent phonemes correct. Nine listeners, blind to diagnoses, rated speech severity using a visual analog scale. Intelligibility was assessed by comparing listeners' orthographic transcriptions of children's responses to target responses.

Results: Measures of speech severity (GFTA-2 standard score, number of unique CAS signs, total CAS signs, and mean severity rating) were significantly correlated with measures of intelligibility (GFTA-2 raw score, percent phonemes correct, and mean intelligibility score). Speech severity and intelligibility did not differ significantly between children with and without LI. Only consonant errors contributed significant variability to speech severity. Consonant errors and stress errors contributed significant variability to intelligibility.

Conclusions: Findings suggest that visual analog scale ratings are a valid and convenient measure of single-word speech severity and that GFTA-2 raw score is an equally convenient measure of single-word intelligibility. The result that consonant errors were by far the major contributor to single-word speech severity and intelligibility in children with CAS, with stress errors also making a small contribution to intelligibility, suggests that consonant accuracy and appropriate lexical stress should be prime therapeutic targets for these children in the context of treatment addressing motor planning/programming, self-monitoring, and self-correcting.

Supplemental Material: https://doi.org/10.23641/asha.19119350

Childhood apraxia of speech (CAS) is a disorder of motor planning and programming that often results in imprecise, inconsistent, and unintelligible speech (American Speech-Language-Hearing Association, 2007). As a sole diagnosis, CAS affects one to two in 1,000 children (Shriberg et al., 1997), but more recent work suggests that it occurs as a comorbidity more frequently in children with complex neurodevelopmental disorders (e.g., Baylis & Shriberg, 2018; Chenausky et al., 2019; Fedorenko et al., 2016; Mei et al., 2018; Raca et al., 2013; Shriberg et al., 2009, 2011, 2019). There is no single abnormal speech characteristic that uniquely identifies CAS, but the American Speech-Language-

Correspondence to Karen V. Chenausky: kvchenausky@mghihp.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

Hearing Association (American Speech-Language-Hearing Association, 2007) put forth three consensus criteria for CAS, while noting that they are neither necessary nor sufficient for diagnosis. They are (a) inconsistent phoneme errors over repeated attempts at the same target, (b) lengthened and disrupted coarticulatory transitions, and (c) inappropriate prosody, especially in the realization of lexical or phrasal stress.

CAS Diagnosis

Diagnosis of the presence and severity of CAS is made by expert clinician judgment, based on a thorough history and observation of a child's performance on a variety of speaking tasks (Strand & McCauley, 2019). During the assessment, clinicians often reference a symptom checklist that contains the abnormal speech characteristics that are associated with CAS and, in particular, note whether a child shows signs consistent with the three consensus criteria to aid diagnosis (Chenausky et al., 2020; Fedorenko et al., 2016; Iuzzini-Seigel et al., 2015; Shriberg et al., 2012, 2017; Strand & McCauley, 2019). A common convention is that some minimum number of signs of CAS (generally any four or five from a list of 10 or 12) must appear in a child's speech across a variety of tasks in order to meet criteria for a CAS diagnosis. However, the relationship of signs of CAS to the underlying deficit is complex and a checklist alone is not sufficient for diagnosis (Strand et al., 2013), so the clinician also provides expert judgment about whether the perceived abnormalities are likely to stem from an underlying impairment of motor programming for speech.

In line with the idea that some signs of CAS are more diagnostic or pathognomonic than others, surveys have identified the speech features that clinicians most often use for differential diagnosis of CAS. For example, Forrest (2003) identified six main characteristics, including inconsistent productions, general oromotor difficulties, and nonspeech groping, which over half of the speech-language pathologist (SLP) respondents used to diagnose CAS. More recently, Randazzo (2019) reported on results of a similar survey, this time with clinicians specializing in CAS. Those clinicians ranked inconsistency of production as the speech feature most indicative of CAS, followed by groping and inappropriate prosody.

Severity of CAS

In addition to diagnosing the presence of CAS, clinicians often determine the severity of CAS. Shriberg and Kwiatkowski (1982) characterize severity as a combination of three underlying concepts: disability (how abnormal the speaker's performance is by some measure), intelligibility (which will be defined below), and handicap (how much the speaker's disability limits effective communication).

Clinicians estimate speech severity for several reasons, including prognostication and therapy planning. Children with more severe CAS might be expected to make less progress over a given length of time and to require more time in therapy to reach a given level of speech proficiency than children with more mild CAS. Severity estimates can therefore inform treatment planning, as severity may affect the length or intensity of treatment required. A valid and convenient measure of severity is also needed for genetic studies in order to understand whether a larger copy number variant of a candidate gene, for example, is associated with more severe CAS.

Measures of CAS Severity

Several measures of severity appear in the literature. Shriberg and Kwiatkowski (1982) proposed percent consonants correct (PCC), calculated from a connected speech sample that is transcribed by a listener, as one metric for severity of pediatric speech sound disorders. PCC and its cousin PCC-R (a revised version of PCC in which typical and atypical consonant distortions have been removed from the index) have been used to assess CAS severity (Iuzzini & Forrest, 2010; Murray et al., 2015). However, these measures carry with them several disadvantages. First, the most severely affected children may not be able to produce connected speech. Second, it is laborious and time consuming to transcribe speech samples from children with speech disorders and to establish adequate interrater reliability on the transcriptions, making transcription-based measures of severity less feasible clinically. Third, PCC can only be calculated for utterances whose target is known, which means that it cannot include unintelligible utterances in spontaneous speech samples. PCC for severely affected children might therefore be inflated by only including the utterances that are intelligible enough to be understood in the count—and children's PCC might actually decrease over time as more partially intelligible utterances replace formerly unintelligible ones and are included in the count.

Another method for assessing severity is by the use of clinician ratings. Shriberg and Kwiatkowski (1982) in fact validated PCC as a measure of severity by comparing it to such ratings. In their study, a group of 55 clinicians and 120 students rated "severity of involvement" by assigning a number between 3 and 7 to 1-min conversational speech samples from 30 children with "delayed speech." PCC accounted for approximately 43% of the variance in these severity ratings. Using an iterative "best-fit" procedure, ranges of PCC values were then established to indicate four ranges of severity: mild (85%–100% PCC), mild to moderate (65%–85%), moderate to severe (50%–65%), and severe (< 50%). The clinicians in Murray et al. (2015) perceptually assigned severity ratings to the speech of the 47 children in their study. Finally, other methods for assessing severity in CAS include summing the total number of different errors across consonant productions (Terband et al., 2019), tallying the number of signs of CAS a child displays on a particular speech task (Iuzzini, 2019), or using scores from standardized articulation tests (Iuzzini, 2019; Murray et al., 2019).

Intelligibility in CAS

Intelligibility, broadly defined as the extent to which a speaker's utterances are understood by a listener (Allison, 2020; Yorkston et al., 1996), is a concept closely related to severity in that the more severe a speech sound disorder, the lower the speaker's intelligibility is likely to be. However, the two are separate concepts (Sussman & Tjaden, 2012). Corroborating this, Shriberg and Kwiatkowski (1982) found a correlation of −.74 between their measures of severity and intelligibility, illustrating both the strong relationship between the two and the fact that they are not identical. Conceptually, the idea of using intelligibility as a proxy for speech severity is appealing because intelligibility can be measured quantitatively using validated techniques such as orthographic transcription of speech (Allison, 2020), which require less specific expertise than phonetic transcription.

Factors Affecting Intelligibility in CAS

Using intelligibility as a proxy for CAS severity carries with it its own set of challenges. A basic one is that intelligibility changes over time even in typically developing children. For example, Hustad et al. (2020) report that the 50th percentile for multiword intelligibility was 40% for 30-month-olds, growing to 78% for 47-month-olds. However, there was also considerable variability between participants (30–40 percentage points at 47 months) and within-age differences were greater than between-age differences. In CAS, too, age has been found to be significantly correlated with intelligibility (McCabe et al., 1998), and thus, understanding how much an impairment in intelligibility may be due to development and how much to a disorder is difficult. It is similarly challenging to separate the amount of change in intelligibility that is due to treatment from that due to maturation.

Different measures of intelligibility may be subject to ceiling or floor effects in some children with CAS. For example, a test on which less severely affected children can achieve close to 100% intelligibility may be too difficult for children who are more severely affected. Furthermore, even among a group of children who achieve the minimum score on a common test of intelligibility, there may be some variation in severity. Children whose CAS is less severe may actually score within the normal range, while, again, there still may be some variability in the group according to severity.

Ratings of children's intelligibility may also be affected by a listener's experience with pediatric speech sound disorders. It has been anecdotally noted and empirically validated that a child's caregivers generally understand a child's speech better than do unfamiliar listeners (Flipsen, 1995; Kwiatkowski & Shriberg, 1992). Similarly, listeners with more experience in pediatric motor speech disorders may find children with CAS more intelligible than listeners with less experience (Allison, 2020). Intelligibility is also greater when a message is highly predictable (Garcia & Cannito, 1996).

Furthermore, not all signs of CAS are expected to affect intelligibility in the same way. For example, the signs *consonant error*, *voicing error*, *nasality error*, *and vowel error* might be expected to significantly reduce intelligibility because they could create meaning-changing errors (e.g., hearing "done" instead of "none"). Other signs, such as *slow rate* or *groping* would not be expected to reduce intelligibility; in fact, speaking slowly might be a compensatory strategy to increase intelligibility (e.g., Blanchet & Snyder, 2010). The effect on intelligibility of the remaining signs is less clear, but still intelligibility might differ between two children who showed the same number, but a different set, of signs of CAS. Finally, the signs that reduce intelligibility might be different from those that are most often used to differentially diagnose CAS—*groping* is commonly used to identify CAS but, since it occurs when a child is not speaking, it is unlikely to affect intelligibility.

The presence of comorbidities may also affect intelligibility. For example, children with language impairment (LI) showed greater movement variability than age-matched peers when producing multisyllabic stimuli (Goffman, 1999, 2004, 2010). Another study, which examined children with cerebral palsy, found that a subgroup of children with comorbid LI and motor speech difficulties was less intelligible than a subgroup with motor speech difficulties but language skills within normal limits (Hustad et al., 2012). Iuzzini-Seigel et al. (2017) investigated inconsistency of production in children with CAS alone and CAS + LI and found that, though the two groups were statistically equivalent on all experimental measures, for some stimuli, there was a trend for the CAS + LI group to show greater inconsistency than the CAS-only group.

Hypotheses

The above issues motivated us to investigate two concepts, speech severity and word intelligibility, at the single-word level in children with CAS. Specifically, we sought to determine the relationship between the two concepts in CAS, the relationship of different signs of CAS to the two concepts, whether the two would be different for more and less experienced listeners, and whether the two differed in children with CAS alone or with CAS + LI. We specifically selected a measure of single-word intelligibility in order to include children of a wide range of severities, some of whom would not be able to complete a sentence production task. Our hypotheses were as follows.

1. Measures of single-word speech severity and intelligibility will be significantly correlated with each other and with chronological age.
2. Children with comorbid LI will have more severe and less intelligible speech than children with CAS alone.
3. Different signs of CAS will have different strengths of association with single-word speech severity and intelligibility.

Method

The study included a group of 30 children with CAS, from two locations, and a group of nine adult listeners, also from two locations. These groups are described below.

Participants

Child Speaker Participants

A total of 30 children with CAS participated in the study. Ten children (three girls) aged 3;9–11;1 (years; months) were seen as part of a previous research study on CAS at Miami University in Ohio (MU). The remaining 20 children (all boys; age range: 4;0–17;0) were seen as part of a previous research study at the University of Nebraska, Lincoln (UNL). Protocols for each study were approved by institutional review boards at MU and UNL, respectively, and parents of all children gave informed, written permission for their child to participate prior to enrollment.

For each data set, multiple criteria were used to verify CAS diagnosis. First, all children had to show at least five of the signs of CAS from the list in Iuzzini-Seigel et al. (2015) across a set of speech tasks, according to judgments of two SLPs with experience in pediatric motor speech disorders. Second, the two SLPs had to agree that participants' performance was consistent with difficulties in motor planning for speech. Third, all children must have scored below the 16th percentile for their age on the Goldman-Fristoe Test of Articulation–Second Edition (GFTA-2, Goldman & Fristoe, 2000). Diagnosis for all children was verified separately for each study by SLPs experienced in pediatric motor speech disorders. Confirmation of CAS diagnosis and scoring of UNL GFTA-2 responses were performed by two SLPs from UNL on the original project for that data (see, e.g., Iuzzini-Seigel et al., 2017). For the MU data, diagnosis was confirmed by a different pair of SLPs (see Chenausky et al., 2020) and GFTA-2 responses were coded by the first and second authors.

In addition to the GFTA-2, children also received standardized tests of language and, in some cases, nonverbal IQ. The eight children from the MU group who were old enough received the Receptive Language Index from the Clinical Evaluation of Language Fundamentals–Fifth Edition (CELF-5; Wiig et al., 2013), while the other two received the Receptive Language subtest of the Mullen Scales of Early Learning (MSEL RL; Mullen, 1995). Children from the UNL group received the full Clinical Evaluation of Language Fundamentals–Fourth Edition (CELF-4; Semel et al., 2003) and the Reynolds Intellectual Assessment Scales (RIAS; Reynolds & Kamphaus, 2003), to assess nonverbal IQ.

Comorbid LI was defined as a standard score on the CELF-5 RLI or the full CELF-4 of 85 or lower (i.e., more than 1 *SD* below the mean), in accordance with guidelines from Wiig et al. (2019) and consistent with Iuzzini-Seigel et al. (2017). Seven of the children from the MU group scored 85 or lower on the CELF-5 RLI; 10 of the children from the UNL group received a score on the CELF-4 of 85 or lower. For the two children who received the MSEL RL, T-scores were first converted into standard scores using a norm score calculator located at https://www.psychometrica.de/normwertrechner_en.html. One child received a standard score of 83 on the MSEL RL (T-score = 33) and was included in the LI group; the other received a standard score of 111 (T-score = 61). Demographic information for all child participants appears in Table 1. Detailed testing information for each participant appears in Supplemental Material S1.

Adult Listener Participants

Adult listeners consisted of five certified SLPs, recruited through the MGH Institute of Health Professions in Boston, MA, and four master's students in the speech-language pathology program at MU. All SLPs had at least 3 years as certified clinicians but were not required to have specific experience with CAS. The protocol under which listener data were collected was approved by the institutional review board at Mass General Brigham, and all adult participants gave written informed consent before participating. The listening experiment was completed remotely by each listener. Study data were collected and managed using REDCap (Research Electronic Data Capture; Harris et al., 2009, 2019), a secure, web-based software platform designed to support data capture for research studies hosted at Mass General Brigham.

Listening Experiment Design

Audio of the GFTA-2 was edited into clips containing only children's responses in a list-like format (i.e.,

Table 1. Demographic information about child speakers.

University	Variable	Age	GFTA-2 standard score	CELF-4 standard score	CELF-5 RLI standard score	MSEL-RL T-score
UNL	*n*	20	20	20		
	mean ± *SD*	9;7 ± 3;1	60.3 ± 15.1	85.5 ± 26.4		
	[min–max]	[4;0–17;0]	[40.0–82.0]	[44.0–133.0]		
MU	*n*	10	10		8	2
	mean ± *SD*	7;3 ± 2;4	59.3 ± 16.2		71.4 ± 16.5	47.0 ± 19.8
	[min–max]	[3;9–11;1]	[40.0–85.0]		[48.0–96.0]	[33, 61]

Note. GFTA-2 = Goldman-Fristoe Test of Articulation–Second Edition; CELF-4 = Clinical Examination of Language Fundamentals–Fourth Edition; CELF-5 RLI = Clinical Examination of Language Fundamentals–Fifth Edition, Receptive Language Index; MSEL-RL: Mullen Scales of Early Learning, Receptive Language subscale; UNL = University of Nebraska, Lincoln; MU = Miami University, Ohio.

experimenter prompts and interjections were removed from the audio file). Listeners were instructed to use their own headphones, to adjust the volume to a comfortable level, and to listen to each word in the clip no more than twice. They were not informed as to the nature of the speech samples (i.e., that they were GFTA-2 responses) or the speakers (i.e., that all were children with a disorder). Instead, they were told that some of the children might have speech sound disorders and some not.

Listeners performed two tasks: (a) orthographically transcribing what they thought they heard and (b) rating each speaker's severity using a visual analog scale (VAS). The VAS consisted of a horizontal line whose ends were labeled "no speech impairment" (corresponding to a value of 0) and "severe speech impairment" (corresponding to a value of 100). Numerical values were not visible to listeners, but the point on the line that a listener selected was automatically translated by REDCap into a score ranging from 0 to 100. Each listener listened to all children, and three children's responses (10%, representing a range of intelligibility levels) were played twice to all participants in order to assess intralistener reliability. All listeners heard GFTA-2 clips in the same order. Figure 1 illustrates the REDCap interface presented to each listener.

Severity and Intelligibility Measures

Several measures of single-word speech severity and intelligibility were derived from the children's GFTA-2 responses. The GFTA-2 was selected not only because it is very commonly administered to children with speech sound disorders, but because it is also appropriate for children who do not speak in full sentences and therefore cannot provide a connected speech sample. This decision comes with limitations, which are discussed in the Limitations and Future Work section. Note that, for clarity, measures have been described as representing either severity or intelligibility. Each measure is described below.

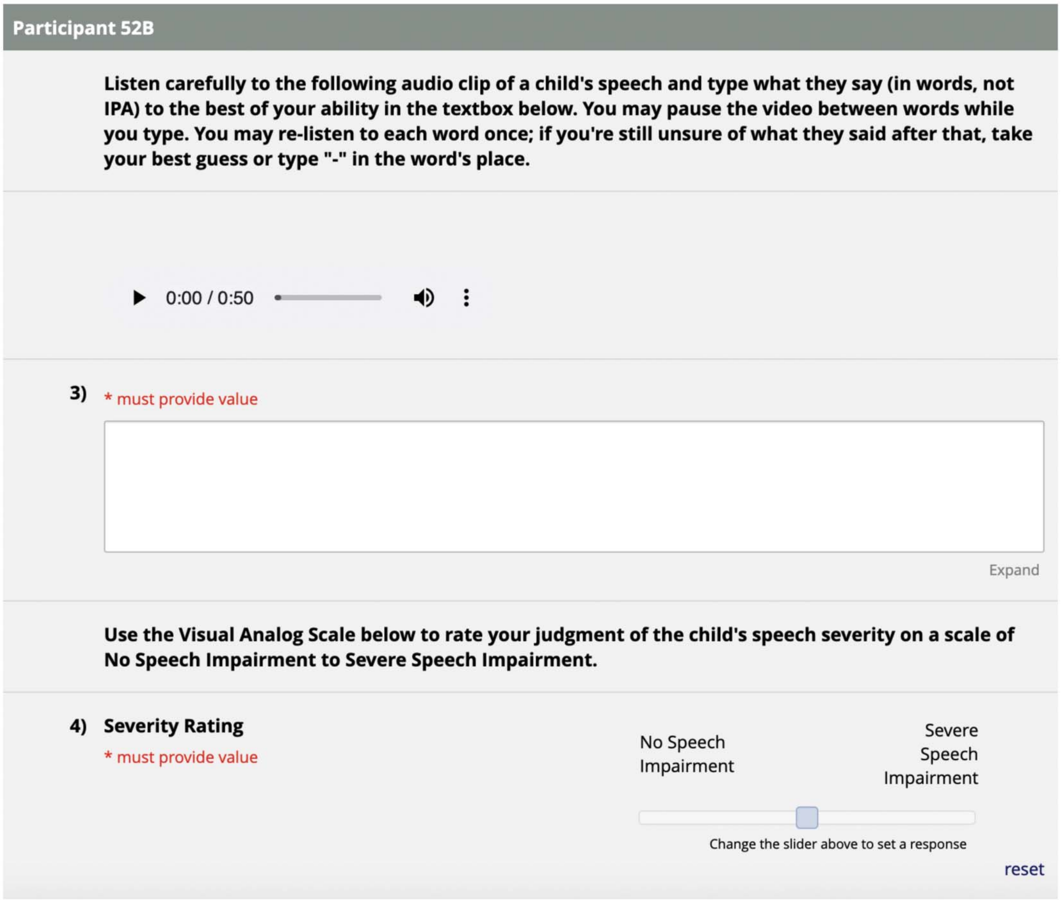
Speech severity measures. Certified SLPs' and master's students' VAS ratings of severity based on single words were averaged to yield *mean severity ratings* for each child. In different analyses, described below, the *mean clinician severity rating*, the *mean student severity rating*, or *overall mean severity rating* was used.

Children's responses on the GFTA-2 were coded by the first two authors for signs of CAS for a previous study using the list of signs from Chenausky et al. (2020). This list is based on that of Iuzzini-Seigel et al. (2015) but differs in some respects. For example, because of the difficulty of differentiating highly distorted consonants from substitutions (i.e., judging how distorted a consonant can be before it is qualifies as a substitution), "consonant errors" included substitutions and omissions as well as distortions. This is consistent with the procedures of other researchers who have included substitutions within the definition of consonant or vowel errors (Shriberg et al., 2009). Operational definitions of the signs appear in the Appendix, and detailed CAS sign information for each participant appears in Supplemental Material S2.

Two measures of single-word severity were extracted from the coded GFTA-2 responses: *total CAS signs,* the total number of signs that appeared over the course of the GFTA-2 for each child; and *unique CAS signs,* the number of different signs of CAS that appeared in each child's GFTA-2. Methods for assuring the reliability of coding GFTA-2 responses for signs of CAS are described below. Finally, GFTA-2 standard scores (GFTA-SS) were also used as a measure of single-word severity.

Word intelligibility measures. Listeners' orthographic transcriptions of children's GFTA-2 responses were compared to the target responses. The number of words correctly transcribed (the number of words in the listeners' transcriptions that matched the target words) was divided by the total number of words on the GFTA-2 to yield a percentage. Scores were averaged to obtain the *mean intelligibility* score for each child. Children's raw scores on the GFTA-2 (GFTA-raw) constituted another measure of

Figure 1. REDCap interface for rating severity and performing orthographic speech transcriptions. IPA = International Phonetic Alphabet.



Participant 52B

Listen carefully to the following audio clip of a child's speech and type what they say (in words, not IPA) to the best of your ability in the textbox below. You may pause the video between words while you type. You may re-listen to each word once; if you're still unsure of what they said after that, take your best guess or type "-" in the word's place.

▶ 0:00 / 0:50

3) * must provide value

Expand

Use the Visual Analog Scale below to rate your judgment of the child's speech severity on a scale of No Speech Impairment to Severe Speech Impairment.

4) Severity Rating
* must provide value

No Speech Impairment | Severe Speech Impairment

Change the slider above to set a response

reset

single-word intelligibility. Finally, the percent phonemes correct (PPC) from each child's GFTA-2 responses, assessed by the first two authors, was the final measure of single-word intelligibility. The coding and reliability procedure for PPC scores from GFTA-2 responses is described below.

Intra- and Interrater Reliability

Intralistener reliability for single-word speech severity and intelligibility. To assess intralistener reliability of the speech severity and word intelligibility measures, a random selection of 10% of the speech files was presented twice to listeners. Two-way mixed intraclass correlation coefficients (ICCs) for average measures and absolute agreement were then calculated (for mean clinician severity ratings, ICC = .932, p = .048; for mean student severity ratings, ICC = .934, p = .068; for mean clinician intelligibility, ICC = .996, p = .001; for mean student intelligibility, ICC = .932, p = .075).

Interlistener reliability for single-word speech severity and intelligibility. Next, we assessed interlistener reliability of the speech severity and intelligibility measures by comparing mean student severity and mean clinician severity values. A two-way mixed ICC for average measures and absolute agreement for single-word speech severity yielded ICC = .999, p = .001. A two-way mixed ICC for average measures and absolute agreement for single-word intelligibility yielded ICC = .984, p = .015.

Perceptual coding reliability. As mentioned, videos of children's GFTA-2 responses were coded for the signs of CAS from Chenausky et al. (2020) to yield two measures of single-word severity. Both the total number of signs that appeared during the GFTA-2 (total CAS signs) and the number of unique signs (unique CAS signs) were tallied for each child. To assess reliability for signs of CAS, 10% of each GFTA-2 video was independently coded. A two-way mixed ICC for single measures and consistency over all signs was used, since one judge's codes were employed in subsequent analyses. The overall ICC was .901, p < .0005. Mean ICC for individual signs was .856, all p < .05. Reliability was not assessed for *unique CAS signs* because this was derived from the total number of signs.

PPC, a measure of word intelligibility, was also calculated from each child's GFTA-2 responses. The same two judges transcribed each GFTA-2 response as part of the coding process; the number of correct phonemes was tallied from this. The average of their scores was used in subsequent analyses. Interjudge reliability was assessed on 10% of the GFTA-2 items using a two-way mixed ICC for average measures and absolute agreement, with ICC = .897, $p < .0005$.

Analytic Strategy

Pearson's correlations were used to assess the degree of association between measures of single-word speech severity, measures of single-word intelligibility, and age, and between CAS signs and measures of single-word speech severity and intelligibility. Paired-samples *t* tests were used to determine whether there were differences between ratings by certified SLPs and master's students. One-way analyses of variance (ANOVAs) were used to determine whether there were differences between scores from children with CAS + LI and children with CAS alone. Finally, hierarchical linear regressions were used to assess how much variance each sign of CAS contributed to measures of single-word speech severity and intelligibility.

Results

Relationship of Single-Word Speech Severity, Intelligibility, and Age

Mean severity rating, total CAS signs, unique CAS signs, and GFTA-SS (measures of single-word speech severity) were entered into Pearson's correlations with PPC, mean intelligibility, and GFTA-raw (measures of single-word intelligibility); and with chronological age. All measures of speech severity were significantly correlated with all measures of intelligibility at the single-word level, with magnitudes of *r* ranging from $|.409|$ to $|.972|$ and all $p \leq .025$. Listener-rated measures of single-word speech severity and intelligibility were negatively correlated with each other. The strongest correlation was between overall mean severity rating and PPC, $r = -.972$, $p < .0005$. GFTA-raw was positively correlated with overall mean severity rating, total CAS signs, and unique CAS signs and was negatively correlated with GFTA-SS. Note that this negative correlation arises because a higher standard score means fewer errors (and thus less impairment), while a higher raw score means more errors (and thus more impairment). Similarly, total CAS signs and unique CAS signs also negatively correlate with intelligibility measures (except for GFTA raw).

Age was significantly correlated with all three single-word measures of intelligibility, with values of *r* ranging from $|.431|$ to $|.639|$, all $p < .017$. Age was also significantly correlated with overall mean severity rating, $r = -.494$, $p = .006$. Age was not significantly correlated with total CAS signs, unique CAS signs, or GFTA-SS. See Table 2 for details. Figure 2 shows scatter plots of variables significantly correlated with age; Figure 3 shows scatter plots for measures of intelligibility and measures of speech severity.

Comparison of Certified SLPs' and Master's Students' Ratings

Paired-samples *t* tests comparing clinician and student single-word VAS severity ratings and single-word intelligibility scores revealed that there was a small but statistically significant difference between the two groups. Mean clinician severity was 25.2 (*SD* = 29.4) and mean student severity was 31.6 (*SD* = 28.0), $p < .0005$. The overall mean severity rating was used in the correlation analysis. However, mean clinician

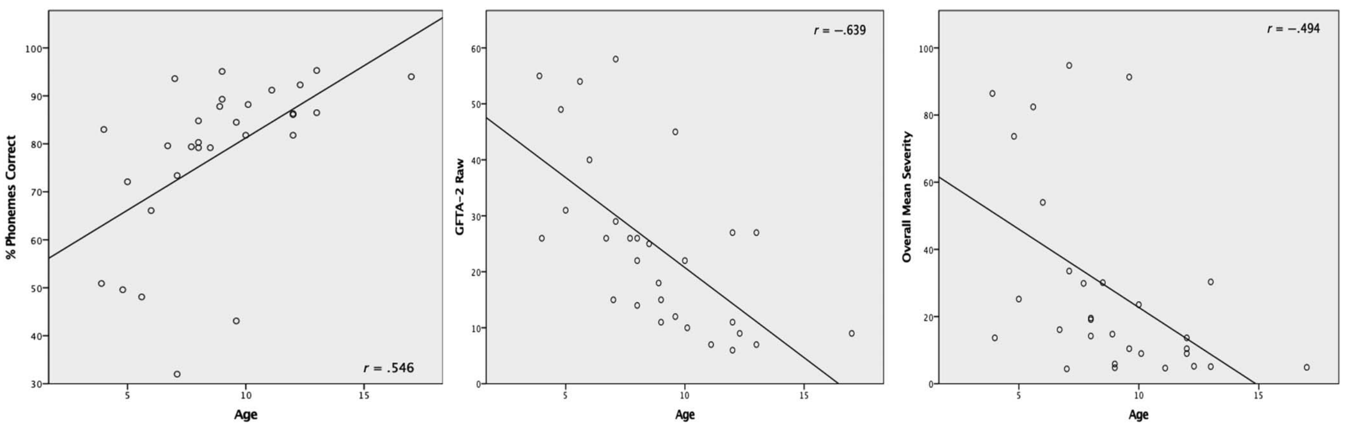
Table 2. Correlations between severity, intelligibility, and age.

		Intelligibility measures			Severity measures			
Measure		**PPC**	**Overall mean intelligibility**	**GFTA-2 raw**	**Overall mean severity**	**Total CAS signs**	**Unique CAS signs**	**GFTA-2 SS**
	Age	.546*	.431*	–.639*	–.494*	–.228	–.446	–.048
Intelligibility measures	**PPC**		.945*	–.926*	–.972*	–.788*	–.582*	.438*
	Overall mean intelligibility			–.868*	–.959*	–.819*	–.541*	–.409*
	GFTA-2 raw				.937*	.749*	.671*	–.544*
Severity measures	**Overall mean severity**					.830*	.609*	–.498*
	Total CAS signs						.567*	–.546*
	Unique CAS signs							–.453*

Note. PPC = percent phonemes correct from GFTA-2 responses; GFTA-2 raw: Goldman-Fristoe Test of Articulation–Second Edition raw score; CAS = childhood apraxia of speech; GFTA-2 SS: Goldman-Fristoe Test of Articulation–Second Edition standard score.

**p* < .05.

Figure 2. Scatter plots of measures that were significantly correlated with age. GFTA-2 Raw: Goldman-Fristoe Test of Articulation–Second Edition raw score.



severity and mean student severity were used in separate regressions to explore whether signs of CAS would relate differently to clinicians' and students' severity ratings.

Mean clinician intelligibility was 83.6 (*SD* = 28.0) and mean student intelligibility was 84.6 (*SD* = 24.2), *p* = .312. Since there was no significant difference between intelligibility scores between clinicians and students, the overall mean single-word intelligibility score was used in all subsequent analyses.

Comparison of Children With CAS Alone and CAS + LI

As described in the Method section, 18 children of the larger group of 30 met criterion for LI. A one-way ANOVA on age between the CAS-only and CAS + LI groups was not significant, *F*(1, 28) = 0.005, *p* = .943. Neither was an ANOVA significant for overall mean speech severity, *F*(1,28) = 0.430, *p* = .517, or overall mean single-word intelligibility, *F*(1, 28) = 1.063, *p* = .311. Details appear in Table 3.

Figure 3. Scatter plots of measures of word intelligibility and measures of speech severity. CAS = childhood apraxia of speech; GFTA-2 Raw = Goldman-Fristoe Test of Articulation–Second Edition raw score; GFTA-SS = GFTA-2 standard scores.

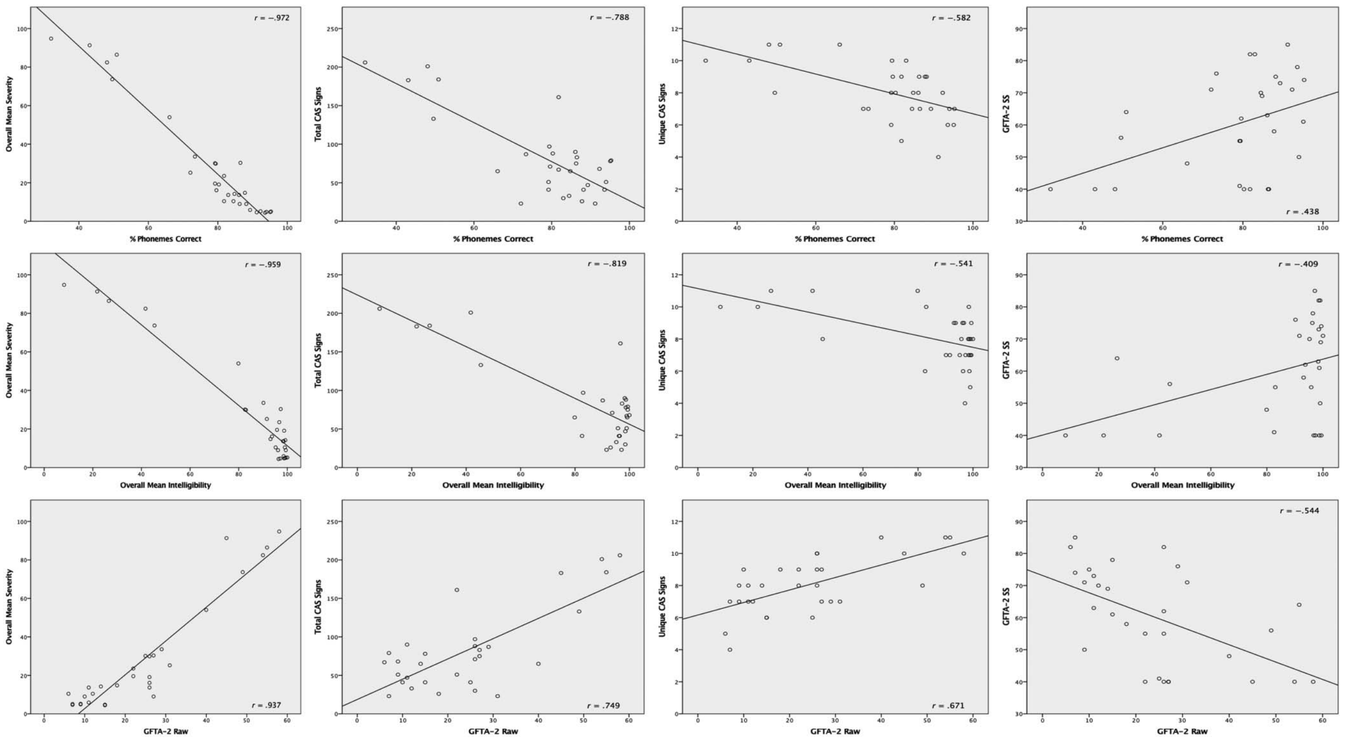


Table 3. Comparison between children with CAS only and CAS + LI.

Group	Age	Overall mean severity	Overall mean intelligibility
CAS-only	8;11 ± 3;10	23.7 ± 21.9	90.1 ± 15.5
n = 12	[4;0–17;0]	[4.4–73.7]	[45.4–99.4]
CAS + LI	8;10 ± 2;6	30.8 ± 32.7	80.0 ± 31.3
n = 18	[4;0–12;4]	[4.7–94.8]	[8.2–100.0]

Note. Figures are reported as mean ± standard deviation, [min–max]. CAS = childhood apraxia of speech; LI = language impairment.

Association Between Signs of CAS and Listener-Rated Single-Word Speech Severity and Intelligibility

To assess the contributions of different CAS signs to mean clinician and mean student severity ratings, we first examined correlations between each sign of CAS and the relevant mean speech severity rating (see Table 4). CAS signs that were significantly correlated with both mean clinician and mean student severity ratings were *consonant error*, *nasality error*, *vowel error*, *syllable segmentation*, *stress error*, *slow rate*, and *addition error*. The sign *groping* was significantly correlated with mean student severity rating but not mean clinician severity rating.

Next, two initial hierarchical regression models were constructed (one for mean clinician severity rating and one for mean student severity rating), in which each significantly correlated sign was entered in order on subsequent steps. Signs that added no significant variability to the model were then removed. Secondary hierarchical regression models were then constructed for each outcome variable, including just the predictor variable(s) from Step 1 that accounted for significant variability in the outcome variable. As in Step 1, each predictor variable was entered on a separate step. The same process was used to construct regression models for overall mean single-word intelligibility. The final models for each outcome variable are summarized in Table 5.

For mean clinician severity rating, the overall initial model was significant, *F*(7, 22) = 18.459, $p < .0005$. No variables were associated with variance inflation factor (VIF) values greater than 6.7, indicating no collinearity. Only the signs *consonant error* and *stress error* accounted for significant variability, so the final regression model contained only these predictor variables. The final model was also significant, *F*(2, 27) = 58.276, $p < .0005$. There was no collinearity (all VIF < 2.0). Both predictors accounted for significant amounts of variance in mean clinician severity rating: *Consonant error* accounted for 76.9% of the variance and *stress error* an additional 4.3%.

For mean student severity rating, the initial overall model was significant, *F*(8, 21) = 12.529, $p < .0005$. There was no collinearity (all VIF < 6.9). Only the sign *consonant error* accounted for significant variability. The final model, with *consonant error* as a unique predictor, was significant, *F*(1, 28) = 70.586, $p < .0005$. *Consonant error* accounted for 71.6% of the variance in mean student severity rating.

Finally, for overall mean single-word intelligibility, the initial model was significant, *F*(8, 21) = 21.719, $p < .0005$. There was no collinearity (all VIF < 7.2). The predictors that accounted for significant variability in overall mean single-word intelligibility were *consonant error*, *stress error*, and *nasality error*. The secondary model, including only those three variables, was also significant, *F*(3, 26) =

Table 4. Correlations between severity, intelligibility, and childhood apraxia of speech (CAS) signs.

CAS sign	Mean clinician severity	Mean student severity	Overall mean severity	Overall mean intelligibility
Consonant error	.877*	.846*	.867*	−.860*
Voicing error	.051	−.008	.027	−.095
Nasality error	.539*	.498*	.525*	−.597*
Vowel error	.543*	.544*	.546*	−.555*
Intrusive schwa	.301	.283	.297	−.344
Syllable segmentation	.707*	.694*	.706*	−.687*
Stress error	.611*	.540*	.585*	−.662*
Slow rate	.667*	.639*	.659*	−.661
Difficulty with transitions	−.012	−.064	−.034	.036
Groping	−.318	−.374*	−.343	.282
Variable errors	.087	.097	.093	−.118
Addition error	.595*	.608*	.604*	−.620*

*p < .05.

Table 5. Regression model statistics.

Regression on mean clinician severity	β	*SE*	*p*	ΔR^2
(Constant)	−21.717	6.470	.002	
Consonant error	0.645	0.081	< .001	.769
Stress error	4.265	1.720	.02	.043
Regression on mean student severity				
(Constant)	−0.069	4.680	n.s.	
Consonant error	0.683	0.081	< .001	.716
Regression on overall mean intelligibility				
(Constant)	129.025	5.738	< .001	
Consonant error	−0.534	0.072	< .001	.739
Stress error	−5.060	1.525	.003	.076

Note. *SE* = standard error; n.s. = not significant.

41.611, $p < .0005$. There was no collinearity (all VIF < 1.8). After this step, *nasality error* did not account for significant variability in overall mean single-word intelligibility and was removed. The final model thus included only *consonant error* and *stress error* and was significant, $F(2, 27) = 59.288$, $p < .0005$. Here, *consonant error* accounted for 73.9% of the variability in overall mean single-word intelligibility and *stress error* another 7.6%.

Discussion

In this article, several results emerged that illuminate the relationship between listener ratings of speech severity, intelligibility, and specific signs of CAS at the single-word level in children with CAS with and without comorbid LI. All of our measures of single-word speech severity and intelligibility were significantly correlated with each other and with age, as predicted. Master's students rated children's speech as slightly more severe than experienced SLPs, as predicted, but there was no difference in single-word intelligibility ratings for the two listener groups. Mean single-word speech severity ratings and single-word intelligibility did not differ significantly between children with CAS + LI and children with CAS alone, in contradiction to predictions. Finally, not all signs of CAS contributed equally to single-word speech severity and intelligibility. Instead, consonant errors accounted for the most variance in these ratings, with stress errors contributing a small additional amount of variance. These findings and related issues will be discussed in turn.

Relationship of Severity, Intelligibility, and Signs of CAS at the Single-Word Level

The expected finding that listener ratings of single-word speech severity were significantly and highly correlated with measures of single-word intelligibility suggests that VAS ratings are one valid way to assess severity in CAS that is also more convenient than, for example, PPC or PCC. Note that the GFTA-2 itself has a severity scale based on its standard score: "average" is within 1.0 *SD* of the mean, "mild" is between 1.0 and 1.5 *SD*s below the mean, "moderate" is between 1.5 and 2.0 *SD*s below the mean, and "severe" is more than 2.0 *SD*s below the mean. Given that mean speech severity ratings were only moderately correlated with GFTA-2 standard scores in our participants ($r = -.498$), clinician-rated severity may be a more accurate measure of severity at the single-word level than categories based on GFTA-2 standard scores.

The significant correlation of age and PPC is consistent with previous work by McCabe et al. (1998). The fact that age was moderately correlated with single-word intelligibility (especially GFTA-raw), but not all measures of speech severity based on single words, lends credence to the idea that severity and intelligibility are indeed separate constructs. Like GFTA-SS, total CAS signs and unique CAS signs were also independent of age. In the case of the GFTA-SS, this is because a standard score explicitly controls for age. The independence of total CAS signs or unique CAS signs from age may be a reflection of the fact that CAS severity is not necessarily age-related.

In terms of the effect of specific signs of CAS on our measures, the finding that consonant errors accounted for the largest amount of variance in single-word speech severity and intelligibility is largely consistent with previous work in dysarthric speakers. Lee et al. (2014) examined a variety of acoustic measures as contributors to word intelligibility in children with cerebral palsy and found that those associated with the articulatory subsystem accounted for the most variance in speech intelligibility (57.9%), while those associated with the laryngeal and velopharyngeal subsystems accounted for only 8.8% and 0.8% of the variance in intelligibility respectively. Rong et al. (2016) also found that deterioration in the articulatory subsystem was the largest contributor to speech intelligibility decline in speakers with amyotrophic lateral sclerosis, accounting for 57.7% of the variance in intelligibility decline, as opposed to 22.7% from the resonatory subsystem, 8.3% from the phonatory subsystem, and 7.2% from the respiratory subsystem. The specific definition of "consonant error" used in this study may have affected the contribution of this sign to the variance in severity and intelligibility. Our definition includes manner and place of articulation errors and substitutions, as well as omissions. Only voicing and nasality errors are covered under separate signs. Our broad definition of consonant errors, therefore, may have magnified its effect on single-word speech severity and intelligibility ratings over other features.

Differences Between Listener Groups

The predicted differences between listener groups' ratings of single-word speech severity and intelligibility were partially upheld, in that student clinicians did rate children's speech as more severe than certified clinicians. While it is statistically significant, the mean difference of 6 percentage points between the two listener groups may or may not be clinically significant (Stipancic et al., 2018). Additional work is required to determine the minimal detectable difference and minimal clinically important difference in speech severity ratings based on single words. Single-word intelligibility, on the other hand, was not different between the two listener groups. Our study is not equipped to determine the source of the difference in severity ratings between certified clinicians and student clinicians for children with CAS. However, Hustad (2007) documented that listener ratings of confidence in what they hear when they listen to dysarthric speakers is only weakly related to their intelligibility scores and speculates that confidence ratings may actually be a proxy for a different phenomenon. Again, investigation of this question is beyond the scope of this study.

Other differences between certified clinicians and students also deserve discussion. Specifically, *groping* was significantly correlated with students' severity ratings but not certified clinicians' severity ratings. On the other hand, *stress errors* significantly predicted certified clinicians' severity ratings, but not those of students. Note, however, that listeners used only audio recordings to make their ratings; they did not see video of participants and they did not rate the presence of signs of CAS. Listeners did not, in fact, know that any of the children had CAS or any other speech sound disorder (though the instructions hinted that this was the case). It is therefore unclear why groping was significantly correlated with students' severity ratings. That said, groping has been shown to be second only to inconsistency as a feature most indicative of CAS over other speech sound disorders in a survey of clinicians who specialize in CAS (Randazzo, 2019). However, while groping may be highly pathognomonic, distinguishing CAS from other speech sound disorders, it does not appear to contribute to single-word intelligibility ratings.

A similar point may be made about vowel errors, which were found in Randazzo (2019) to be the third most distinguishing feature for CAS and have received a great deal of attention from other researchers (Lenoci et al., 2020). Some authors have advocated for vowel errors to be a central diagnostic feature of CAS (e.g., Jacks et al., 2013). In this study, vowel errors were significantly correlated with single-word speech severity and intelligibility and may very well function as a pathognomonic feature of CAS but did not contribute significant variance to either the severity or intelligibility ratings. This may have been due to the fact that the GFTA-2 provides fewer opportunities for vowel errors to arise than for consonant errors to arise. On the other hand, it may be that, like groping, vowel errors are highly indicative of CAS yet do not compromise single-word intelligibility to a great degree. Errors such as syllable segregation or slow rate may have the same quality of being more closely associated with CAS than with other speech sound disorders, yet not necessarily indicative of single-word severity per se or greatly affecting intelligibility.

Finally, it is important to address the very high, but nonsignificant, ICC values for student intrarater reliability for severity and intelligibility. According to Liljequist et al. (2019), ICC values above .9 indicate excellent reliability, regardless of the associated *p* value. What the *p* value indicates is whether the two-way mixed model (which was used here and models bias as a fixed factor) differs significantly from the one-way model (which assumes no bias). A *p* value greater than .05 indicates that the two-way mixed and one-way models are not significantly different. This means that in fact there was no bias in the repeated student ratings. In other words, while the clinicians may have benefited slightly from repeated presentations of the same speaker, students did not appear to have done so.

Effects of LI Comorbidity on Single-Word Severity and Intelligibility

Another finding was that, contrary to our hypothesis, children with CAS + LI did not have more severe or less intelligible speech at the single-word level, on average, than children with CAS alone. This may have been due to our use of a single-word task, rather than a connected-speech task. If listener participants had been asked to rate spontaneous speech samples for severity and transcribe them orthographically, severity might well have been higher and intelligibility lower because connected speech is more challenging for children with CAS than single-word production (Iuzzini-Seigel et al., 2017) and may involve unfamiliar topics. Still, if the contributions of different conditions to speech severity or intelligibility are viewed as additive, as implied by the findings of Hustad et al. (2012) that children with cerebral palsy and LI were less intelligible than children with cerebral palsy alone, then our finding is surprising. On the other hand, our finding is consistent with those of Iuzzini-Seigel et al. (2017), who examined inconsistency of production on different stimulus types by (among other groups) children with CAS and children with both CAS and LI. Specifically, they found that phonemic inconsistency, calculated in part on GFTA-2 responses, was not significantly different between the CAS and CAS + LI groups. Note, however, that some of the participants in that study also participated in the current study, which may account for the convergent findings.

Work by Pennington (2006) presents a theoretical view of how different comorbidities may interact. Instead of viewing each identified condition (diagnosis) as contributing an independent amount of speech severity or intelligibility, Pennington's model views each diagnosis as resulting from the interaction of multiple factors that may confer either risk or protection. Because some risk factors are shared across diagnoses, high levels of comorbidity are to be expected. The specific risk and protective factors that characterize each child's profile alter that child's development and produce the behavioral signs that result in a specific diagnosis, but the different factors do not necessarily have an additive effect on severity. The different effects of risk factors on diagnosis/comorbidities and severity of those conditions therefore require more research. Altogether, a model of how developmental disorders such as CAS and LI interact will necessarily be complex and must take into account etiological, neural, and cognitive processes that are shared between different disorders. In any case, however, the current results argue against a one-to-one mapping between different risk factors and thus the idea that each disorder produces its own independent deficit and has an additive effect on single-word speech severity.

Limitations and Future Work

This study has several limitations. One is the relatively small number of both listeners and speakers and, potentially, the large amount of heterogeneity in those speakers. Thus, the findings apply to single-word speech severity and intelligibility within a cohort. More research must be done to determine how well they generalize to other cohorts, to changes in single-word speech severity and intelligibility for particular children over time, and to measures of severity and intelligibility from spontaneous speech. The current findings may apply better to tests of single-word intelligibility than to measures of speech severity or intelligibility that are derived from spontaneous conversation or natural language samples.

Despite its convenience, there are disadvantages to using the GFTA-2 as a source of stimuli. First, its emphasis on consonants may have increased the effect of that sign of CAS on severity and intelligibility ratings over what would be found in spontaneous speech. Though they were not informed that the speech samples came from the GFTA-2, it is possible that our clinician raters might have been familiar with the items in this very commonly administered test and rated intelligibility higher than they might have otherwise. This would be consistent with previous findings by, for example, Garcia and Cannito (1996) that predictability of sentence stimuli was associated with significantly higher intelligibility scores in adults with flaccid dysarthria (though note that these authors also found that listening to audio only, rather than audio and video together, was associated with lower intelligibility scores).

A related concern is how some signs of CAS are linked to each other. For example, once *syllable segmentation* is identified, *slow rate* must also be. This has been noted by other researchers (Strand et al., 2013) and is not unique to our study but is a consequence of how CAS signs are defined. Similarly, signs such as *intrusive schwa*, *additions*, *consonant error*, and *syllable segregation* might all plausibly be manifestations of *difficulty with coarticulatory transitions*, meaning that the latter sign would necessarily co-occur with any of the former. Further work must be done to understand the relationship of different error types to each other and to the underlying construct of difficulty with motor programming and planning.

Another limitation is the design of the listening experiment itself. Specifically, all listeners heard the clips in the same order, which may have induced a learning effect, at least in the clinicians. Also, as mentioned, familiarity with the GFTA-2 and the fact that it is a closed set of words, rather than open-ended like conversation, may have increased intelligibility ratings. Thus, further work must be done to understand the relationship of intelligibility and severity as derived from tests of single-word production to similar measures derived from spontaneous speech. For example, a more highly controlled set of words, presented in random order to a larger group of listeners and balanced for syllable structures and phonemes, could be used. The findings for this stimulus set could then be correlated with GFTA scores and measures derived from spontaneous speech to assess the validity of using measures derived from the GFTA going forward.

Clinical Implications

Two main clinical implications emerge from this work. First, VAS-rated estimates perform well as other measures of speech severity based on single-word samples and are much easier to obtain than measures of speech severity such as tallying the total number of CAS signs in a child's speech during the GFTA or calculating PCC from a connected speech sample — especially for children who do not produce much connected speech. In addition, the GFTA raw score also performs well as a measure of single-word intelligibility. While no measure of severity or intelligibility is recommended as a way to differentially diagnose CAS, VAS ratings and GFTA raw score can function as convenient, easy-to-obtain summaries of how severely affected or intelligible children with CAS are, at the single-word level.

The second important clinical implication relates to the finding that consonant error was by far the greatest contributor to speech severity and intelligibility, at least in a test of single-word production. Further work must be done to identify contributors to speech severity and intelligibility in sentence-intelligibility tests or connected speech, but the current findings suggest that focusing treatment on accurate consonant production and, to a lesser extent on lexical stress, may go a long way toward improving the ability of children with CAS and related disorders to produce intelligible speech. However, although attention to these two signs of CAS may have the strongest effect on intelligibility, other aspects of CAS should still be addressed. In particular, improving children's motor programming and planning, as well as their ability to self-monitor and correct their own speech, remain vital goals for children with CAS.

Acknowledgments

The authors would like to thank the children who participated in the studies leading to this manuscript and their families for allowing them to participate. This work was based on the second author's master's thesis and was supported by K99 DC 017490 (awarded to K. V. C.), by P50 DC 018006 (PI: H. Tager-Flusberg, supporting K. V. C. and J. R. G.), and by K24 DC 016312 (awarded to J. R. G.).

References

Allison, K. (2020). Measuring speech intelligibility in children with motor speech disorders. *Perspectives of the ASHA Special Interest Groups, 5*(4), 809–820. https://doi.org/10.1044/2020_PERSP-19-00110

American Speech-Language-Hearing Association. (2007) *Childhood apraxia of speech* [Technical report]. http://www.asha.org/policy

Baylis, A., & Shriberg, L. (2018). Estimates of the prevalence of speech and motor disorders in youth with 22q11.2 deletion syndrome. *American Journal of Speech-Language Pathology, 28*(1), 53–82. https://doi.org/10.1044/2018_AJSLP-18-0037

Blanchet, P., & Snyder, G. (2010). Speech rate treatments for individuals with dysarthria: A tutorial. *Perceptual and Motor Skills, 110*(3), 965–982. https://doi.org/10.2466/pms.110.3.965-982

Chenausky, K., Brignell, A., Morgan, A., Gagné, D., Norton, A., Tager-Flusberg, H., Schlaug, G., Shield, A., & Green, J. (2020). Factor analysis of signs of childhood apraxia of speech. *Journal of Communication Disorders, 87,* 106033. https://doi.org/10.1016/j.jcomdis.2020.106033

Chenausky, K., Brignell, A., Morgan, A., & Tager-Flusberg, H. (2019). Motor speech impairment predicts expressive language in minimally verbal, but not low verbal, individuals with autism spectrum disorder. *Autism and Developmental Language Impairment, 4,* 1–12. https://doi.org/10.1177/2396941519856333

Fedorenko, E., Morgan, A., Murray, E., Cardinaux, A., Mei, C., Tager-Flusberg, H., Fisher, S. E., & Kanwisher, N. (2016). A highly penetrant form of childhood apraxia of speech due to deletion of 16p11.2. *European Journal of Human Genetics, 24*(2), 302–306. https://doi.org/10.1038/ejhg.2015.149

Flipsen, P., Jr. (1995). Speaker–listener familiarity: Parents as judges of delayed speech intelligibility. *Journal of Communication Disorders, 28*(1), 3–19. https://doi.org/10.1016/00219924(94)00015-R

Forrest, K. (2003). Diagnostic criteria of developmental apraxia of speech used by clinical speech-language pathologists. *American Journal of Speech-Language Pathology, 12*(3), 376–380. https://doi.org/10.1044/1058-0360(2003/083)

Garcia, J., & Cannito, M. (1996). Influence of verbal and nonverbal contexts on the sentence intelligibility of a speaker with dysarthria. *Journal of Speech and Hearing Research, 39*(4), 750–760. https://doi.org/10.1044/jshr.3904.750

Goffman, L. (1999). Prosodic influences on speech production in children with specific language impairment and speech deficits. *Journal of Speech, Language, and Hearing Research, 42*(6), 1499–1517. https://doi.org/10.1044/jslhr.4206.1499

Goffman, L. (2004). Kinematic differentiation of prosodic categories in normal and disordered language development. *Journal of Speech, Language, and Hearing Research, 47*(5), 1088–1102. https://doi.org/10.1044/1092-4388(2004/081)

Goffman, L. (2010). Dynamic interaction of motor and language factors in normal and disordered development. In B. Maassen & P. van Lieshout (Eds.), *Speech motor control: New developments in basic and applied research*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199235797.003.0008

Goldman, R., & Fristoe, M. (2000). *Goldman–Fristoe Test of Articulation* (2nd ed.). Pearson. https://doi.org/10.1037/t15098-000

Harris, P., Taylor, R., Thielke, R., Minor, B., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, F., Kirby, J., Duda, S. N., & REDCap Consortium. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics, 95,* 103208. https://doi.org/10.1016/j.jbi.2019.103208

Harris, P., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. (2009). Research electronic data capture (REDCap): A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics, 42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Hustad, K. (2007). Effects of speech stimuli and dysarthria severity on intelligibility scores and listener confidence ratings for speakers with cerebral palsy. *Folia Phoniatrica et Logopaedica, 59*(6), 306–317. https://doi.org/10.1159/000108337

Hustad, K., Mahr, T., Natzke, P., & Rathouz, P. (2020). Development of speech intelligibility between 30 and 47 months in typically developing children: A cross-sectional study of growth. *Journal of Speech, Language, and Hearing Research, 63*(6), 1675–1687. https://doi.org/10.1044/2020_JSLHR-20-00008

Hustad, K., Schueler, B., Schultz, L., & DuHadway, C. (2012). Intelligibility of 4-year-old children with and without cerebral palsy. *Journal of Speech, Language, and Hearing Research, 55*(4), 1177–1189. https://doi.org/10.1044/1092-4388(2011/11-0083)

Iuzzini, J. (2019). Motor performance in children with childhood apraxia of speech and speech sound disorders. *Journal of Speech, Language, and Hearing Research, 62*(9), 3220–3233. https://doi.org/10.1044/2019_JSLHR-S-18-0380

Iuzzini, J., & Forrest, K. (2010). Evaluation of a combined treatment approach for childhood apraxia of speech. *Clinical Linguistics & Phonetics, 24*(4–5), 335–345. https://doi.org/10.3109/02699200903581083

Iuzzini-Seigel, J., Hogan, T., & Green, J. (2017). Speech inconsistency in children with childhood apraxia of speech, language impairment, and speech delay: Depends on the stimuli. *Journal of Speech, Language, and Hearing Research, 60*(5), 1194–1210. https://doi.org/10.1044/2016_JSLHR-S-15-0184

Iuzzini-Seigel, J., Hogan, T., Guarino, A., & Green, J. R. (2015). Reliance on auditory feedback in children with childhood apraxia of speech. *Journal of Communication Disorders, 54,* 32–42. https://doi.org/10.1016/j.jcomdis.2015.01.002

Jacks, A., Marquardt, T., & Davis, B. (2013). Vowel production in childhood and acquired apraxia of speech. In M. Ball & F. Gibbons (Eds.), *Handbook of vowels and vowel disorders*. Psychology Press. https://doi.org/10.4324/9780203103890.ch12

Kwiatkowski, J., & Shriberg, L. (1992). Intelligibility assessment in developmental phonological disorders: Accuracy of caregiver gloss. *Journal of Speech and Hearing Research, 35*(5), 1095–1104. https://doi.org/10.1044/jshr.3505.1095

Lee, J., Hustad, K., & Weismer, G. (2014). Predicting speech intelligibility with a multiple speech subsystems approach in children with cerebral palsy. *Journal of Speech, Language, and Hearing Research, 57*(5), 1666–1678. https://doi.org/10.1044/2014_JSLHR-S-13-0292

Lenoci, G., Celata, C., Ricci, I., Chilosi, A., & Barone, V. (2020). Vowel variability and contrast in childhood apraxia of speech: Acoustics and articulation. *Clinical Linguistics & Phonetics, 35*(11), 1011–1035. https://doi.org/10.1080/02699206.2020.1853811

Liljequist, D., Elfving, B., & Skavberg Roldsen, K. (2019). Intraclass correlation – A discussion and demonstration of basic features. *PLOS ONE, 14*(7), Article e0219854. https://doi.org/10.1371/journal.pone.0219854

McCabe, P., Rosenthal, J., & McLeod, S. (1998). Features of developmental dyspraxia in the general speech-impaired population. *Clinical Linguistics & Phonetics, 12*(2), 105–126. https://doi.org/10.3109/02699209808985216

Mei, C., Fedorenko, E., Amor, D., Boys, A., Hoeflin, C., Carew, P., Burgess, T., Fisher, S., & Morgan, A. (2018). Deep phenotyping of speech and language skills in individuals with 16p11.2 deletion. *European Journal of Human Genetics, 26*(5), 676–686. https://doi.org/10.1038/s41431-018-0102-x

Mullen, E. M. (1995). *Mullen scales of early learning*. AGS.

Murray, E., Iuzzini-Seigel, J., Maas, E., Terband, E., & Ballard, K. (2019). Differential diagnosis of childhood apraxia of speech compared to other speech sound disorders: A systematic review. *American Journal of Speech-Language Pathology, 30*(1), 279–300. https://doi.org/10.1044/2020_AJSLP-20-00063

Murray, E., McCabe, P., Heard, R., & Ballard, K. (2015). Differential diagnosis of children with suspected childhood apraxia of speech. *Journal of Speech, Language, and Hearing Research, 58*(1), 43–60. https://doi.org/10.1044/2014_JSLHR-S-12-0358

Pennington, B. (2006). From single to multiple deficit models of developmental disorders. *Cognition, 101*(2), 385–413. https://doi.org/10.1016/j.cognition.2006.04.008

Raca, G., Baas, B., Kirmani, S., Laffin, J., Jackson, C., Strand, E., Jakielski, K., & Shriberg, L. (2013). Childhood apraxia of speech (CAS) in two patients with 16p11.2 microdeletion syndrome. *European Journal of Human Genetics, 21*(4), 455–459. https://doi.org/10.1038/ejhg.2012.165

Randazzo, M. (2019). A survey of clinicians with specialization in childhood apraxia of speech. *American Journal of Speech-Language Pathology, 28*(4), 1659–1672. https://doi.org/10.1044/2019_AJSLP-19-0034

Reynolds, C., & Kamphaus, R. (2003). *Reynolds intellectual assessment scales*. Psychological Assessment Resources.

Rong, P., Yunusova, Y., Wang, J., Zinman, L., Pattee, G., Berry, J., Perry, B., & Green, J. (2016). Predicting speech intelligibility decline in amyotrophic lateral sclerosis based on the deterioration of individual speech subsystems. *PLOS ONE, 11*(5), Article e0154971. https://doi.org/10.1371/journal.pone.0154971

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals* (4th ed.). Pearson.

Shriberg, L. (2008). Childhood apraxia of speech (CAS) in neuro-developmental and idiopathic contexts. In R. Sock, S. Fuchs & Y. Laprie (Eds.), *Proceedings of the 8th International Seminar on Speech Production* (pp. 193–196). Université March Bloch.

Shriberg, L., Aram, D., & Kwiatkowski, J. (1997). Developmental apraxia of speech. *Journal of Speech, Language, and Hearing Research, 40*(2), 273–285. https://doi.org/10.1044/jslhr.4002.273

Shriberg, L., & Kwiatkowski, J. (1982). Phonological disorders II: A conceptual framework for management. *Journal of Speech and Hearing Disorders, 47*(3), 242–256. https://doi.org/10.1044/jshd.4703.242

Shriberg, L., Lohmeier, H., Strand, E., & Jakielski, K. (2012). Encoding, memory, and transcoding deficits in childhood apraxia of speech. *Clinical Linguistics & Phonetics, 26*(5), 445–482. https://doi.org/10.3109/02699206.2012.655841

Shriberg, L., Potter, N., & Strand, E. (2009). *Childhood apraxia of speech in children and adolescents with galactosemia*. American Speech-Language-Hearing Association.

Shriberg, L., Potter, N., & Strand, E. (2011). Prevalence and phenotype of childhood apraxia of speech in youth with galactosemia. *Journal of Speech, Language, and Hearing Research, 54*(2), 487–519. https://doi.org/10.1044/1092-4388(2010/10-0068)

Shriberg, L., Strand, E., Fourakis, M., Jakielski, K., Hall, S., Karlsson, H., Mabie, H., McSweeny, J., Tilkens, C., & Wilson, D. (2017). A diagnostic marker to discriminate childhood apraxia of speech from speech delay: I. Development and description of the pause marker. *Journal of Speech, Language, and Hearing Research, 60*(4), S1096–S1117. https://doi.org/10.1044/2016_JSLHR-S-16-0148

Shriberg, L., Strand, E., Jakielski, K., & Mabie, H. (2019). Estimates of the prevalence of speech and motor speech disorders in persons with complex neurodevelopmental disorders. *Clinical Linguistics & Phonetics, 33*(8), 707–736. https://doi.org/10.1080/0s699206.2019.1595732

Stipancic, K., Yunusova, Y., Berry, J., & Green, J. (2018). Minimally detectable change and minimal clinically important difference of a decline in sentence intelligibility and speaking rate for individuals with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research, 61*(11), 2757–2771. https://doi.org/10.1044/2018_JSLHR-S-17-0366

Strand, E. A., & McCauley, R. J. (2019). *Dynamic Evaluation of Motor Speech Skill (DEMSS) manual*. Brookes.

Strand, E. A., McCauley, R. J., Weigand, S. D., Stoeckel, R. E., & Baas, B. S. (2013). A motor speech assessment for children with severe speech disorders: Reliability and validity evidence. *Journal of Speech, Language, and Hearing Research, 56*(2), 505–520. https://doi.org/10.1044/1092-4388(2012/12-0094)

Sussman, J. E., & Tjaden, K. (2012). Perceptual measures of speech from individuals with Parkinson's disease and multiple sclerosis: Intelligibility and beyond. *Journal of Speech, Language, and Hearing Research, 55*(4), 1208–1219. https://doi.org/10.1044/1092-4388(2011/11-0048)

Terband, H., Namasivayam, A., Maas, E., van Brenk, F., Mailend, M.-L., Diepeveen, S., van Lieshout, P., & Maassen, B. (2019). Assessment of childhood apraxia of speech: A review/tutorial of objective measurement techniques. *Journal of Speech, Language, and Hearing Research, 62*(8S), 2999–3032. https://doi.org/10.1044/2019_JSLHR-S-CSMC7-19-0214

Wiig, E., Semel, E., & Secord, W. (2013). *Clinical Evaluation of Language Fundamentals–Fourth Edition (CELF-4)*. Pearson.

Wiig, E., Semel, E., & Secord, W. (2019). *Determining the severity of a language disorder*. Pearson.

Yorkston, K. M., Strand, E. A., & Kennedy, M. R. T. (1996). Comprehensibility of dysarthric speech: Implications for assessment and treatment planning. *American Journal of Speech-Language Pathology, 5*(1), 55–66. https://doi.org/10.1044/1058-0360.0501.55

Appendix

Operational Definitions of Signs of CAS

Definitions are taken from Chenausky et al. (2020), adapted from Iuzzini-Seigel et al. (2015).

1. **Consonant error**: A consonant production error in which a speech sound is recognizable as a specific phoneme but is not produced exactly correctly (e.g., an /s/ that is produced with lateralization or dentalization). Also includes substitutions (e.g., [to] for "so") and omissions (e.g., [ba] for "bob"). Not scored if the only consonant error is voicing or nasality. Thus, consonant errors include manner/place distortions or substitutions, and omissions.
2. **Voicing error**: A sound is produced as its voicing cognate (e.g., a /p/ that is produced as a [b]). In addition, this could also describe productions which appear to be in between voicing categories (e.g., blurring of voicing boundaries). Note that glottal stop is considered neither voiced nor voiceless, so substitution of a glottal stop for another consonant does not trigger this error.
3. **Nasality error**: Sounds either hyponasal (not enough airflow out of nose/"stuffy") OR hypernasal (too much airflow out of nose for nonnasal phonemes such as plosives). Nasality errors can also occur if an oral stop is substituted for a nasal (e.g., [do] for "no"), if a nasal is substituted for an oral stop (e.g., [mi] for "bee"), or if a vowel in a word with no nasal consonant is heavily nasalized.
4. **Vowel error:** A vowel production error in which the vowel is substituted for another phoneme OR in which the vowel is recognizable as a specific phoneme but is not produced exactly correctly (e.g., it is not a prototypical production but may sound like it is in between two vowels). It is not considered an error if the vowel is substituted with another phoneme that is consistent with an adult-like model or a regional accent (e.g., /hɑtdɑg/, /hɑtdɔg/).
5. **Intrusive schwa** (e.g., in clusters): A schwa is added between consonants. For example, it may be inserted in between the consonants in a cluster (e.g., /blu/ becomes /bəlu/). This NOT considered a "vowel error." Intrusive schwa may also occur before an initial consonant (e.g., [əbʌni] for "bunny") or adjacent to a vowel (e.g., [noə] for "no").
6. **Syllable segregation**: Brief or lengthy pause between syllables within a word which is not appropriate.
7. **Stress error**: An error in which the appropriate stress is not produced correctly. For example: conDUCT and CONduct have different stress patterns. It is considered an error if the stress is inappropriately equalized across syllables or placed on the wrong syllable. Addition of syllables (as in [dædədi] for "daddy") or deletion of syllables (as in [tɛfon] for "telephone") also count as stress errors, since they change the metrical structure of the word.
8. **Slow rate**: Speech rate is not typical. It is slower during production of part (e.g., zzziiiiiiper/zipper) or the whole word (e.g., tooommmmaaatoooo/tomato). Syllable segregation also triggers the "slow rate" error.
9. **Difficulty with coarticulation**: Initiation of utterance or initial speech sound may be difficult for child to produce and may sound lengthened, uncoordinated, or excessively effortful. Also, child may evidence lengthened or disrupted coarticulatory gestures or movement transitions from one sound to the next. For example, heavily prevoiced stops or words with a glottal stop inserted at the beginning fall into this category.
10. **Groping**: Prevocalic (silent) articulatory searching prior to onset of phonation, possibly in an effort to improve the accuracy of the production. Video is needed to assess this feature.
11. **Variable errors**: The same target is produced with different errors each time. Note that if a child produces an errored token once and a correct version once, this does not count as a variable error. The child must produce at least two distinct errored versions in order to trigger this error.
12. **Additions** (of phonemes other than schwa): The token contains phonemes or syllables that are not in the target. For example, [mɑmbi] for "mommy" would contain [b] as an addition (and would also trigger a "difficulty with coarticulation" error). Addition of syllable(s) also triggers the "stress" error.